

Magellan: Web Based Analysis Of Cancer Genomics Data

Data Analysis & Statistical Tools SIG
November 5, 2004

Chris Kingsley
ckingsley@cc.ucsf.edu
Jain Lab
UCSF Cancer Center

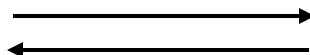
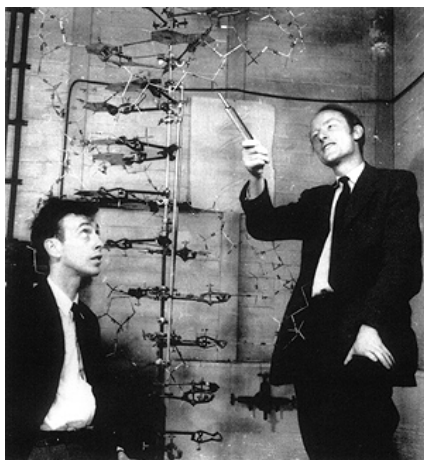
Motivation

- ◆ Many researchers at the UCSF CC were moving toward high throughput methodologies in a number of different areas
 - Array based mRNA expression, CGH, proteomics, methylomics
- ◆ How are most biologists dealing with these megavariable data sets?
 - Various algorithms/applications have been appearing
 - Excel macros, SAM, Spotfire, Bioconductor, custom apps
 - Range of functionality / usability / intimidation
 - Biostatistics gurus
- ◆ How do we deliver the functionality while exploiting the domain knowledge of the biologists?

Analysis of High Throughput Biological Data

Motivation

- ◆ Many researchers at the UCSF CC were moving toward high throughput methodologies in a number of different areas
 - Array based mRNA expression, CGH, proteomics, methylomics
- ◆ How are most biologists dealing with these megavariable data sets?
 - Various algorithms/applications have been appearing
 - Excel macros, SAM, Spotfire, Bioconductor, custom apps
 - Range of functionality / usability / intimidation
 - Biostatistics gurus
- ◆ How do we deliver the functionality while exploiting the domain knowledge of the biologists?

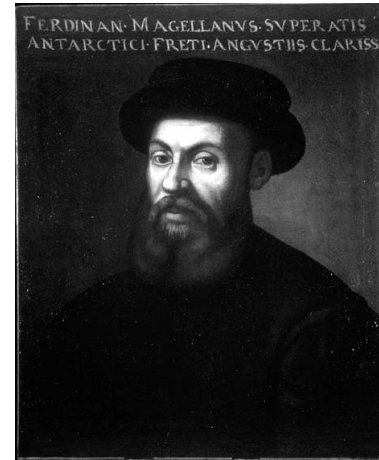


Goals

- ◆ Give biologists themselves the ability to perform analyses, such that their domain knowledge is used. Build an intuitive web based system with general application.
 - Multiple, user specified data types of arbitrary dimension
- ◆ Allow the use of biological annotation information
 - Many different quantitative / qualitative annotations can be linked to data
- ◆ Deploy analytical methods in a modular fashion for ease of extensibility
- ◆ Give users the ability to perform operations on their data prior to analysis
 - Sub selection, projection, etc.

Implementation - Magellan

- ◆ Web based
- ◆ Client – Server Model with a centralized Oracle 8 database
- ◆ Dynamic page content generated with Java/JSP
- ◆ Analytical methods generated in C or R thus far



Generality and Expandability is Key

- ◆ Represent data and annotations abstractly to handle as much information as possible
 - Data is derived from samples
 - Annotations describe variables of a data type
- ◆ Do not impose a nomenclature on users but insist on consistency
 - Imposing identifier nomenclature is a double edged sword – I'd rather use carrots (like databases of curated annotations) than sticks
- ◆ Do not impose particular file formats on data uploads.
- ◆ Try to minimize the pain of interfacing analytical applications
 - Provide functionality in Java Classes with a well documented API
- ◆ Provide a number of generalized operations on data that can be combined
 - Projection, sub selection, import, export, visualizations, etc.

Don't restrict the analytical tools that can be interfaced

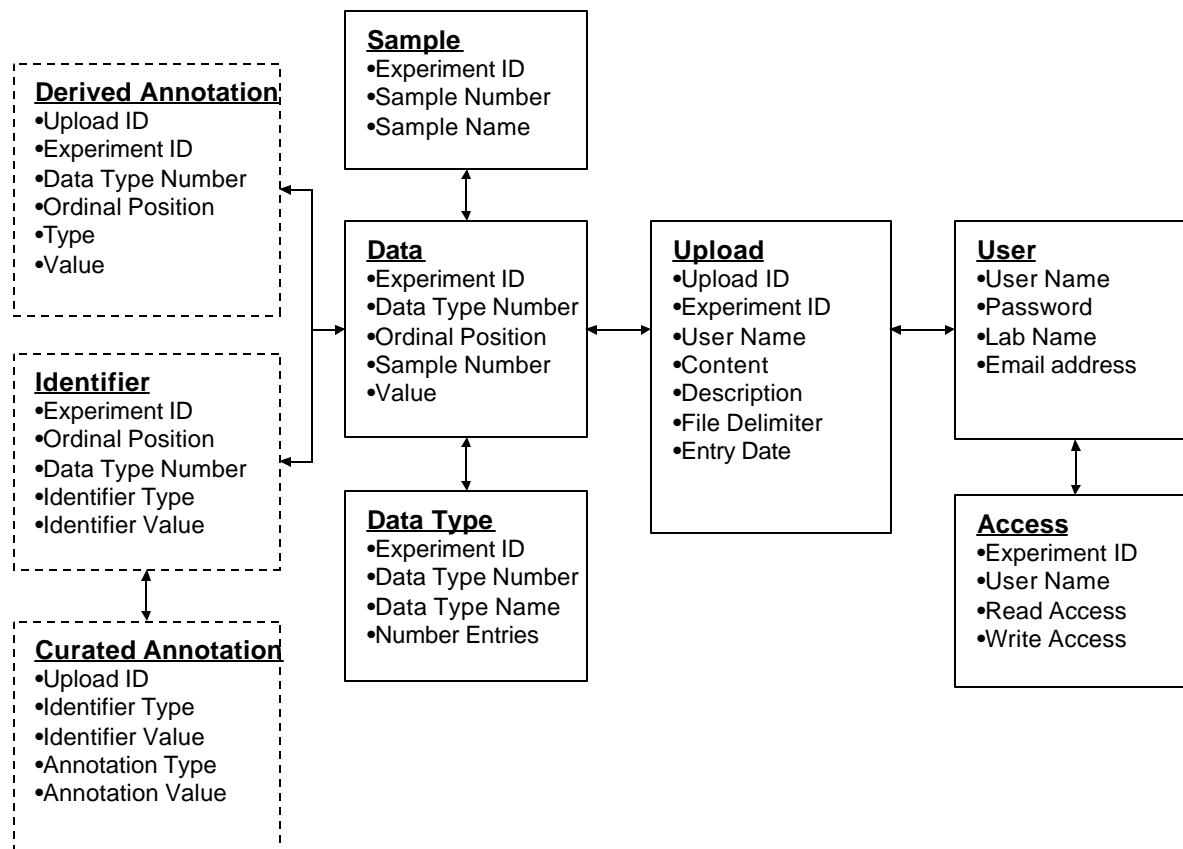
- ◆ Use command line access to non-Java apps, and use flat files for data transfer.
 - Other file formats can be generated by overriding Java methods
 - In the case of R, a common data structure was adopted
 - A Java method dynamically generates R code to load the flat file contents into that structure.
- ◆ Processes are forked off and the system waits for the appearance of the result file.
 - Computation done server side, but should be scalable.
 - Some results can be automatically stored as derived annotations

Make no assumptions as to the type / content of data and annotations (EAV)

- ◆ Information stored as type – value pairs of strings

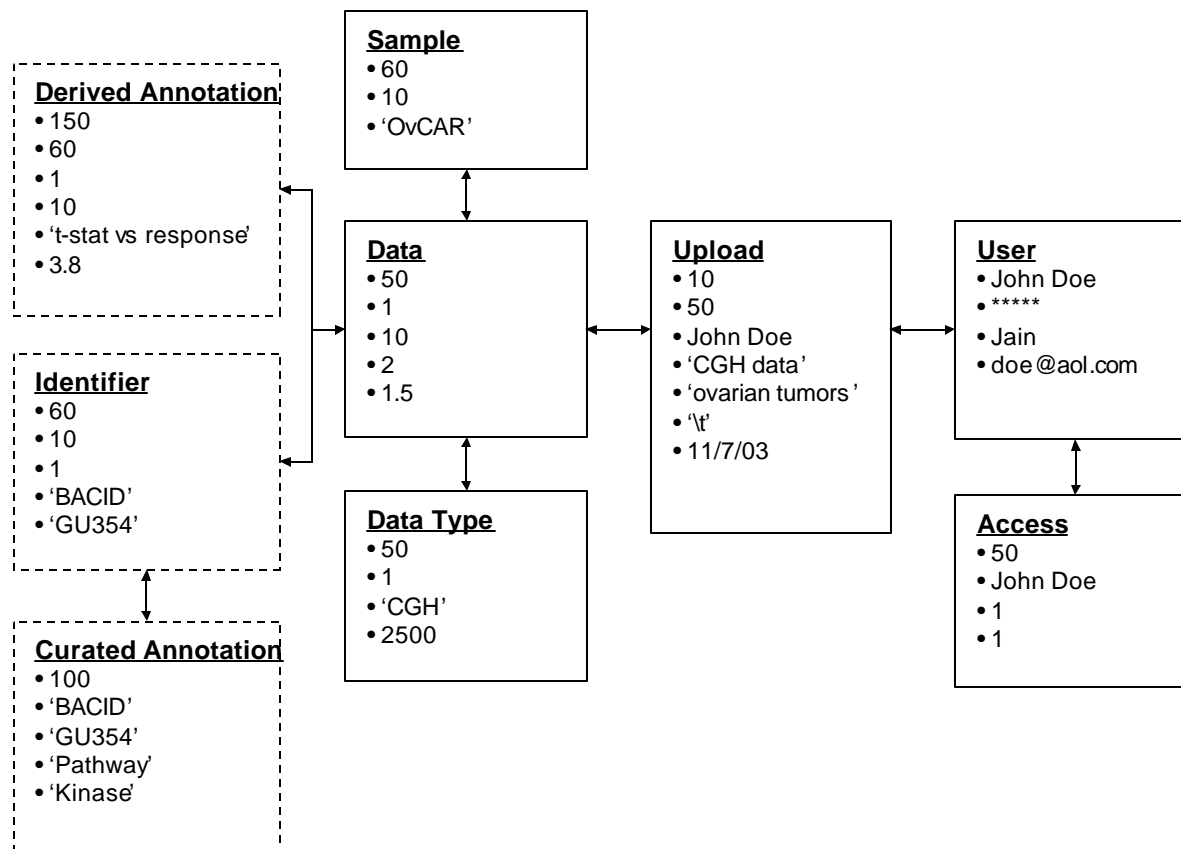
Make no assumptions as to the type / content of data and annotations (EAV)

◆ Information stored as type – value pairs of strings



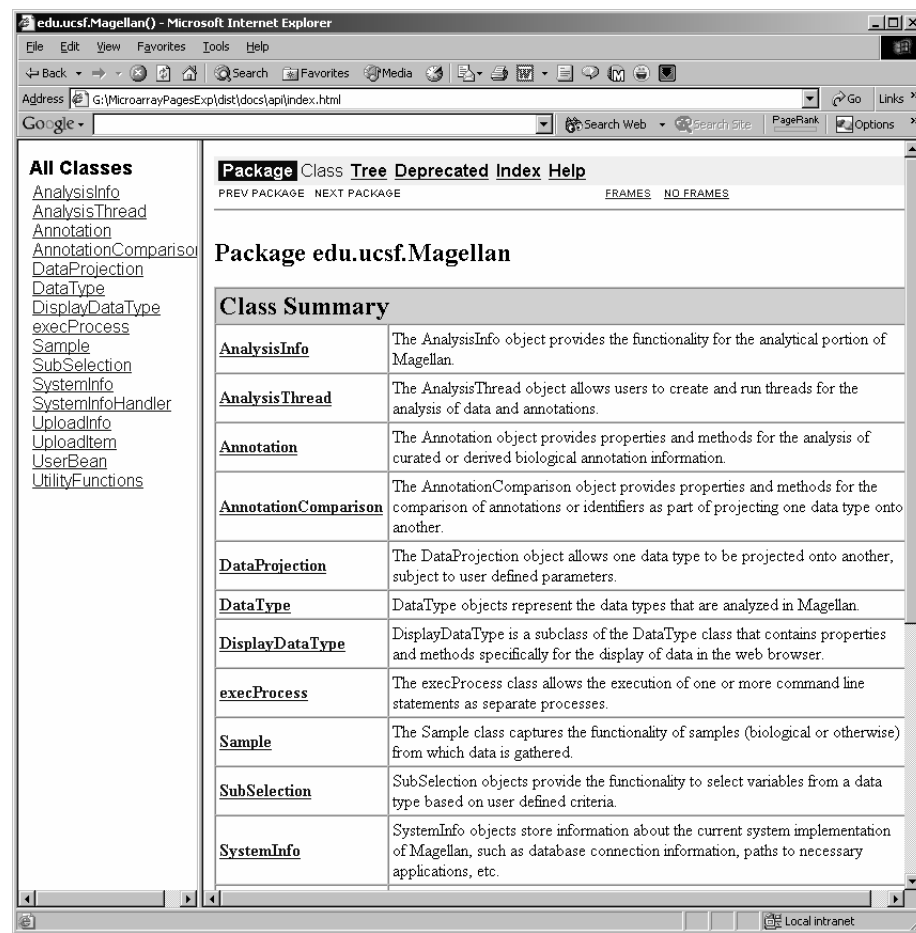
Make no assumptions as to the type / content of data and annotations (EAV)

- ◆ Information stored as type – value pairs of strings



Data Representation

- ◆ All information represented by compiled Java Classes accessible from JSP pages
 - Methods allow developers to specify analytical parameters, generate data files, fork processes, etc.

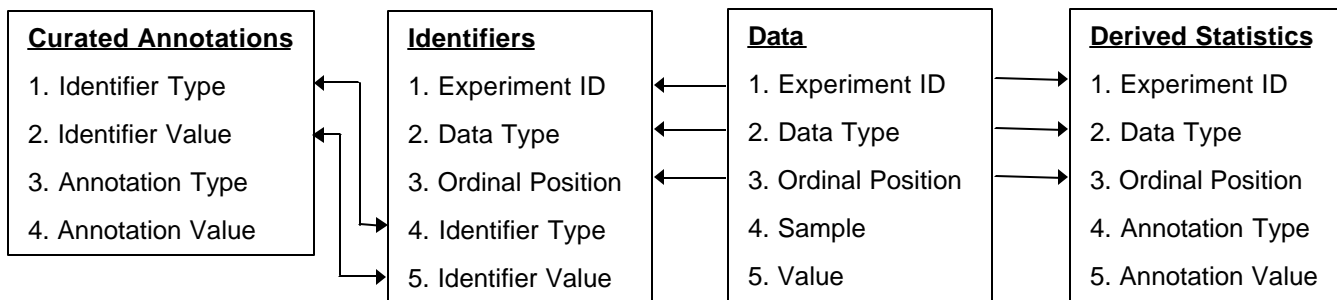


Annotations describe variables of a data type

- ◆ Chromosomal position of genes, pathway designation, correlation with outcome, etc.
- ◆ Annotations can be used by certain algorithms / data operations
- ◆ Data and annotations are linked in two ways, depending on the type of annotations
 - Curated annotations – Applicable to many Data Sets. Linked through textual ‘identifiers’ such as genbank ID’s
 - Derived annotations – Specific to one data set. Linked by row number

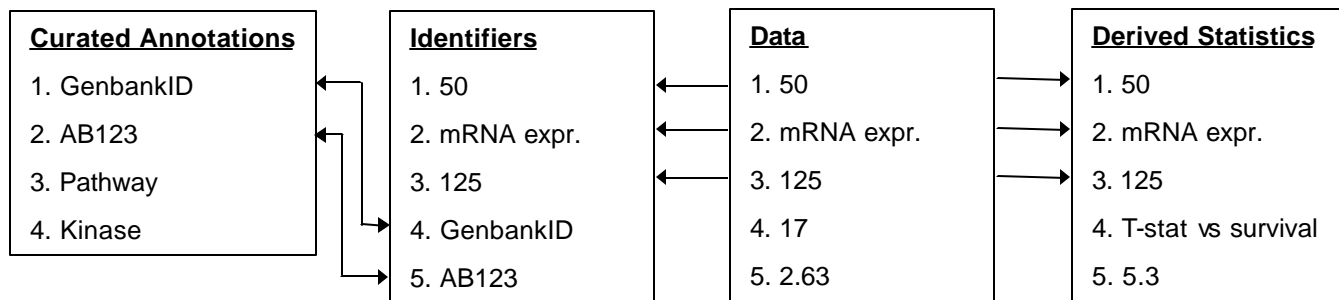
Annotations describe variables of a data type

- ◆ Chromosomal position of genes, pathway designation, correlation with outcome, etc.
- ◆ Annotations can be used by certain algorithms / data operations
- ◆ Data and annotations are linked in two ways, depending on the type of annotations
 - Curated annotations – Applicable to many Data Sets. Linked through textual ‘identifiers’ such as genbank ID’s
 - Derived annotations – Specific to one data set. Linked by row number



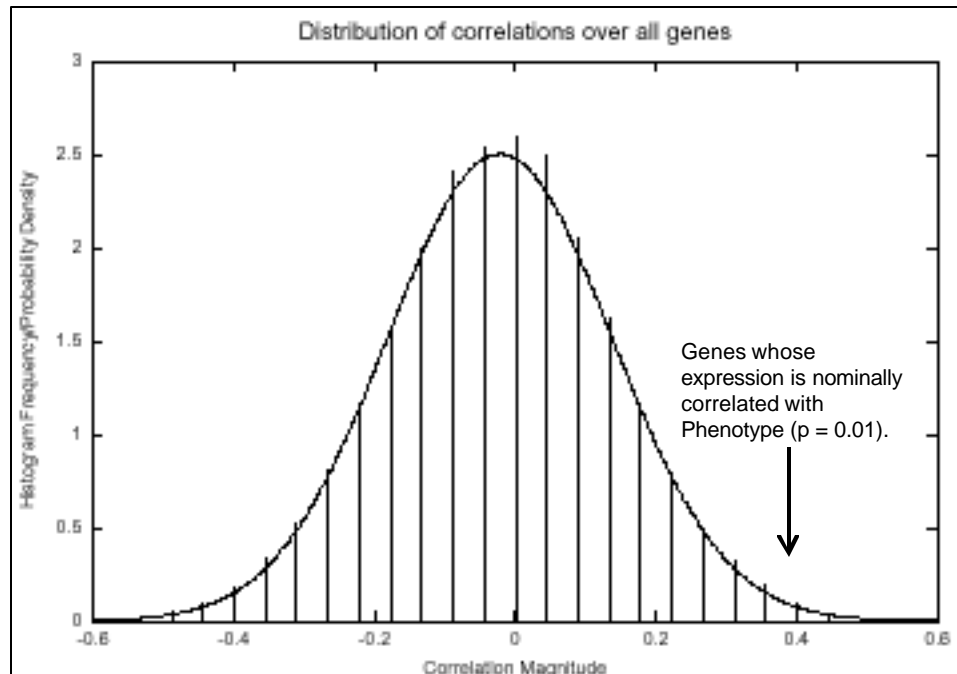
Annotations describe variables of a data type

- ◆ Chromosomal position of genes, pathway designation, correlation with outcome, etc.
- ◆ Annotations can be used by certain algorithms / data operations
- ◆ Data and annotations are linked in two ways, depending on the type of annotations
 - Curated annotations – Applicable to many Data Sets. Linked through textual ‘identifiers’ such as genbank ID’s
 - Derived annotations – Specific to one data set. Linked by row number



Data Sub Selection

- ◆ Data sets can be sub selected based on quantitative or qualitative annotations
 - Allows the creation of biologically meaningful subsets
 - Set size reduction can reduce the effects of multiple comparisons.



No imposed file formats - The User defines the type and location of the uploaded information

File Contents - Microsoft Internet Explorer

Address: http://localhost:8080/MicroarrayAnalysisDev/DatabasePages/upload/displayFileContents.jsp

File Contents

The hyperlinks at the top and left sides are used to define the contents of the uploaded file.

Use the links to indicate the location of the sample names, data types, and identifiers (in that order)

Displaying the first 25 lines of the uploaded file:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	Clone	1526	447	871	1859	475	1451	1590	1834	1837	2596	2422	567	1023	1169	1462	1813	1924	2358	2421
2	total_jeam_M	0.441306	0.362742	0.356262	0.052131	0.374275	0.150007	0.108748	0.216574	0.220828	0.256157	0.538961	0.284649	0.083302	0.220726	0.25076	0.326797	0.207309	0.250639	0.312
3	Vysis-lptel_218C			0.661695			0.016213			0.176894	-0.11759									
4	RP11-82d16				1.132657	0.16411							0.293276	0.338845	0.285223	0.082917	0.237744	0.198813	0.080	
5	RP11-62m23				-1.18816	0.010061							0.019212	0.217973	0.134676	-0.20472	0.026358	0.029331	0.012	
6	RP11-60j11	0.004486	0.054871	0.535917	0.520128	0.082918	0.003431	0.078858	0.420953	0.174284	0.014725	0.057116	0.092182	0.087688	0.271395	0.107102	0.150437	0.150554	0.093477	0.027
7	RP11-111005					0.11517							0.125992	0.34969	0.134025	0.036235		0.254413	0.124	
8	RP11-51b04		0.041865	0.348081	0.495698	0.24701	0.010862	0.040078	0.424975	0.169226	0.078369	0.046962	0.151625	0.103699	0.288857	0.107088	0.139612	0.20668	0.064279	0.002
9	RP11-199a01	0.148007	0.013832	0.376795	0.295353	0.121814	0.093495	0.10104	0.367935	0.054577	0.074944	0.005205	0.088202	0.141597	0.387985	0.071967	0.053743	0.073612	0.221385	0.099
10	RP11-013305	0.055751	0.093381	0.363191	0.344595	0.061126	0.107851	0.180199	0.296682	0.122712	0.087681	0.075909	0.138286	0.068218	0.332126	0.245814	0.067895	0.055084	0.209057	0.135
11	RP11-188b07	0.005774	0.079238	0.331199	0.383198	0.09564	0.019322	0.071228	0.408752	-0.00085	0.050096	0.079217	0.119764	0.033383	0.245461	0.150542	0.165397	0.080044	0.146412	0.058
12	RP11-178m15	0.009790	0.166951	0.411394	0.414240	0.061964	0.012491	0.098061	0.373728	0.048245	0.054494	0.017453	0.056573	0.015539	0.251687	0.167469	0.087150	0.170197	0.173899	0.086

Done Local intranet

File Description - Microsoft Internet Explorer

Address: http://localhost:8080/MicroarrayAnalysisDev/DatabasePages/upload/enterFileInfo.jsp?format=row&num=2

File Description Row 2

Select the type of information that this row contains:
Refer back to the numbered hyperlinks in the table to determine row and column numbers

☐ This row contains Sample Names:
Sample names begin in column and end in column
☒ Sample names are continuous in this row
☐ Sample names are discontinuous - skip column(s) between sample names

☐ This File contains data from a single sample, named:

☒ This row determines the boundary of a data type:
Data begins in row and ends in row
Enter the name of the data type (each name can be used only once):
☐ CGH
☐ Expression
☐ Other

☒ The data for each sample of this data type is in the same column as the sample name

☐ Samples of this data type begin in column and end in column
☒ Samples of this data type are continuous in this row
☐ Samples of this data type are discontinuous - skip column(s) between columns of data

Done Local intranet

Information is previewed prior to upload

Upload - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Mail

Address http://localhost:8080/MicroarrayAnalysisDev/DatabasePages/upload/loadDB.jsp?action=preview Go Links

Google Search Web PageRank 301 blocked AutoFill Options

Upload Preview

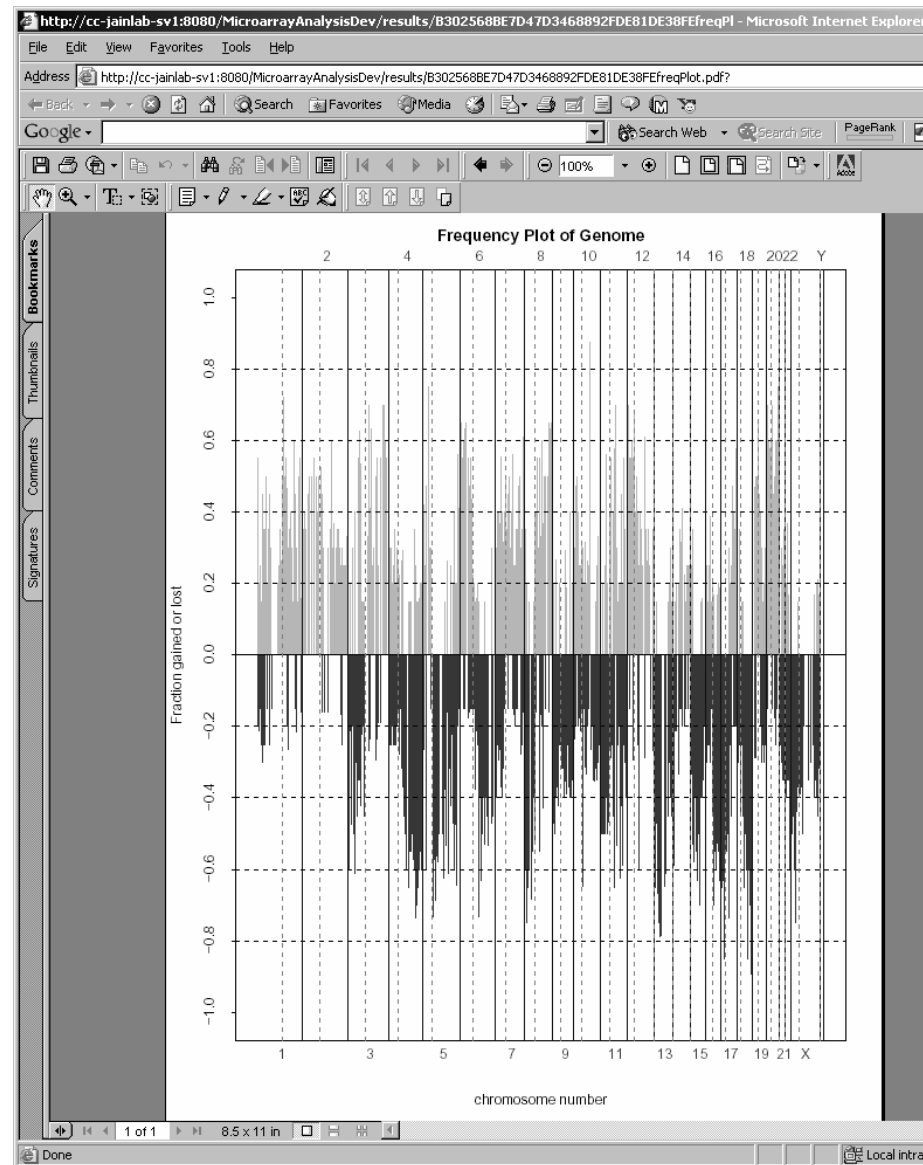
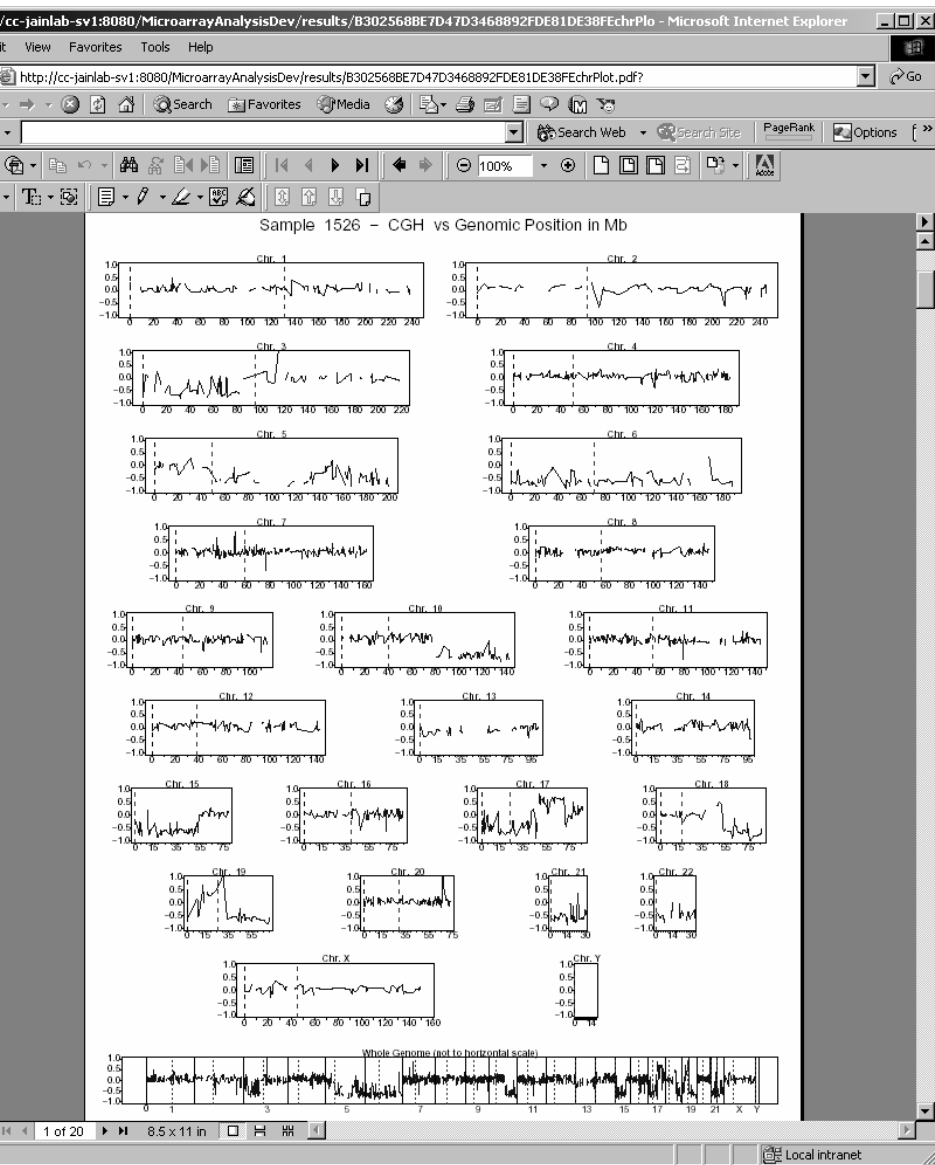
You have specified the following information for the current upload:
 Sample names are indicated in red
 Identifiers are headed in green
 Annotations are headed in blue

Data Type: CGH
 Number of data points per sample: 2364
 Displaying the first 10 lines of data type CGH

Bac Id	1526	447	871	1859	475	1451	1590	1834	1837	2596	2422	567	1023	1169	1462	1813	1924	2358	2421
total_hum_M	0.441306	0.362742	0.356262	0.052131	0.374275	0.150007	0.108748	0.216374	0.220828	0.256157	0.538961	0.284849	0.083302	0.220726	0.25076	0.326797	0.207309	0.250639	0.312258
Vysis- Iptel_218C			0.661695			0.016213			0.176894	-0.11759									
RP11-82d16				1.132657	0.16411								0.293276	0.358845	0.285223	0.082917	0.237744	0.198813	0.080337
RP11-62m23				-1.18816	0.010061								0.019212	0.217973	0.134676	-0.20472	0.026358	0.029331	0.012944
RP11-60j11	0.004486	0.054871	0.535917	0.520128	0.082918	0.003431	0.078858	0.420953	0.174284	0.014725	0.057116	0.092182	0.087688	0.271395	0.107102	0.150437	0.150554	0.093477	0.027621
RP11-111005					0.11517								0.125992	0.34969	0.134025	0.036235		0.264413	0.124811
RP11-51b04		0.041865	0.348081	0.495698	0.24701	0.010862	0.040078	0.424975	0.169226	0.078369	0.046962	0.151625	0.103699	0.288857	0.107088	0.139612	-0.20668	0.064279	0.002265
RP11-199o01	0.148007	0.013832	0.376795	0.295353	0.121814	0.093495	0.10104	0.367935	0.054577	0.074944	0.005205	0.088202	0.141597	0.387985	0.071967	0.053743	0.073612	0.221385	0.099388
RP11-813J05	0.055751	0.093381	0.262101	0.244505	0.061126	0.107851	0.180199	0.208687	0.122712	0.087681	0.075909	0.138386	0.068718	0.332126	0.245814	0.087805	0.055084	0.209057	0.13597

Done Local intranet

Other Analytical Functions



Other Analytical Functions

cc-jainlab-sv1:8080/MicroarrayAnalysisExp/results/070124D1D32342EE86CE18FE3289A406corr.txt - Microsoft Internet Explorer

View Favorites Tools Help

http://cc-jainlab-sv1:8080/MicroarrayAnalysisExp/results/070124D1D32342EE86CE18FE3289A406corr.txt? Go Links »

Search Web Search Site PageRank »

```

: 2596 1837 1834 1590 1451 475 1859 871 447 1526
: 2108 2421 2358 1924 1813 1462 1169 1023 567 2422
Expression row Expression identifier class row class identifier t-test abs(t-t
HG1612-HT1612_at 1 3.3400 3.3400 2 0.9990
M28879_at 1 3.7975 3.7975 1 0.8880
M88108_at 1 3.3518 3.3518 2 0.9990
U07802_at 1 3.3578 3.3578 1 0.9980
U10550_at 1 3.3565 3.3565 1 0.9980
U14747_at 1 3.3885 3.3885 1 0.9980
U19718_at 1 4.0693 4.0693 2 0.6950
U21049_at 1 3.7679 3.7679 1 0.8970
U40369_rna1_at 1 5.4986 5.4986 1 0.0370
U56637_at 1 3.3302 3.3302 1 0.9990
U94333_at 1 3.8133 3.8133 2 0.8840
X66141_at 1 3.9574 3.9574 1 0.7910
U19713_s_at 1 3.4098 3.4098 1 0.9960
HG627-HT5097_s_at 1 3.6540 3.6540 2 0.9550
M16707_rna1_s_at 1 3.3763 3.3763 1 0.9980
Z50115_s_at 1 3.5878 3.5878 1 0.9700
Correlations with p-values less than 1.0 are printed

Maximum absolute correlation observed in 1000 permutations:
3.3241
3.3521
3.3886
3.4075
3.4353
3.4512
3.4610
3.4801
3.4913
3.4984
3.5090

```

Local intranet

Microsoft Excel - 070124D1D32342EE86CE18FE3289A406corr.txt

Arial 10 B I U

File Edit View Insert Format Tools Data Window Help Acrobat

H25 =

	A	B	C	D	E	F	G	H
1	Group 1: 2596 1837 1834 1590 1451 475 1859 871 447 1526							
2	Group 2: 2108 2421 2358 1924 1813 1462 1169 1023 567 2422							
3	Expression row	Expression identifier	class row	class identifier	t-test	abs(t-test)	higher in group	p-value
4	804	HG1612-HT1612_at	1		3.34	3.34	2	0.999
5	1895	M28879_at	1		3.7975	3.7975	1	0.888
6	2324	M88108_at	1		3.3518	3.3518	2	0.999
7	2677	U07802_at	1		3.3578	3.3578	1	0.998
8	2735	U10550_at	1		3.3565	3.3565	1	0.998
9	2795	U14747_at	1		3.3885	3.3885	1	0.998
10	2884	U19718_at	1		4.0693	4.0693	2	0.695
11	2910	U21049_at	1		3.7679	3.7679	1	0.897
12	3163	U40369_rna1_at	1		5.4986	5.4986	1	0.037
13	3413	U56637_at	1		3.3302	3.3302	1	0.999
14	3978	U94333_at	1		3.8133	3.8133	2	0.884
15	4429	X66141_at	1		3.9574	3.9574	1	0.791
16	5683	U19713_s_at	1		3.4098	3.4098	1	0.996
17	5899	HG627-HT5097_s_at	1		3.654	3.654	2	0.955
18	6195	M16707_rna1_s_at	1		3.3763	3.3763	1	0.998
19	6524	Z50115_s_at	1		3.5878	3.5878	1	0.97
20	Only correlations with p-values less than 1.0 are printed							
21								
22	Maximum absolute correlation observed in 1000 permutations:							
23	1	3.3241						
24	2	3.3521						
25	3	3.3886						
26	4	3.4075						
27	5	3.4353						
28	6	3.4512						
29	7	3.461						
30	8	3.4801						
31	9	3.4913						
32	10	3.4984						
33	11	3.509						

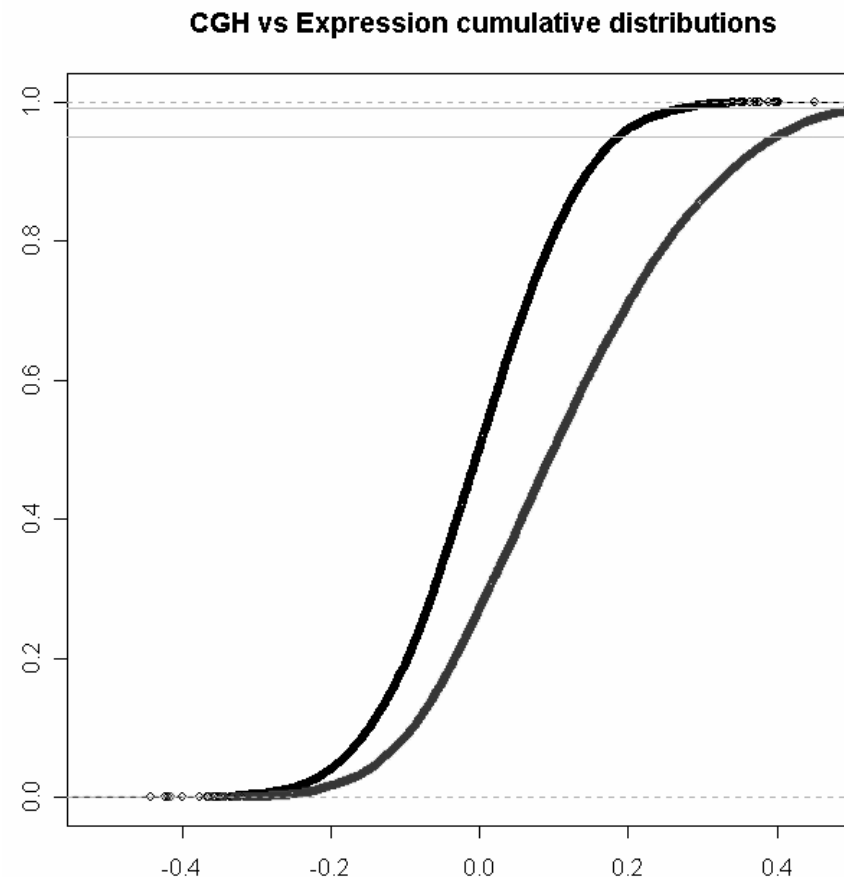
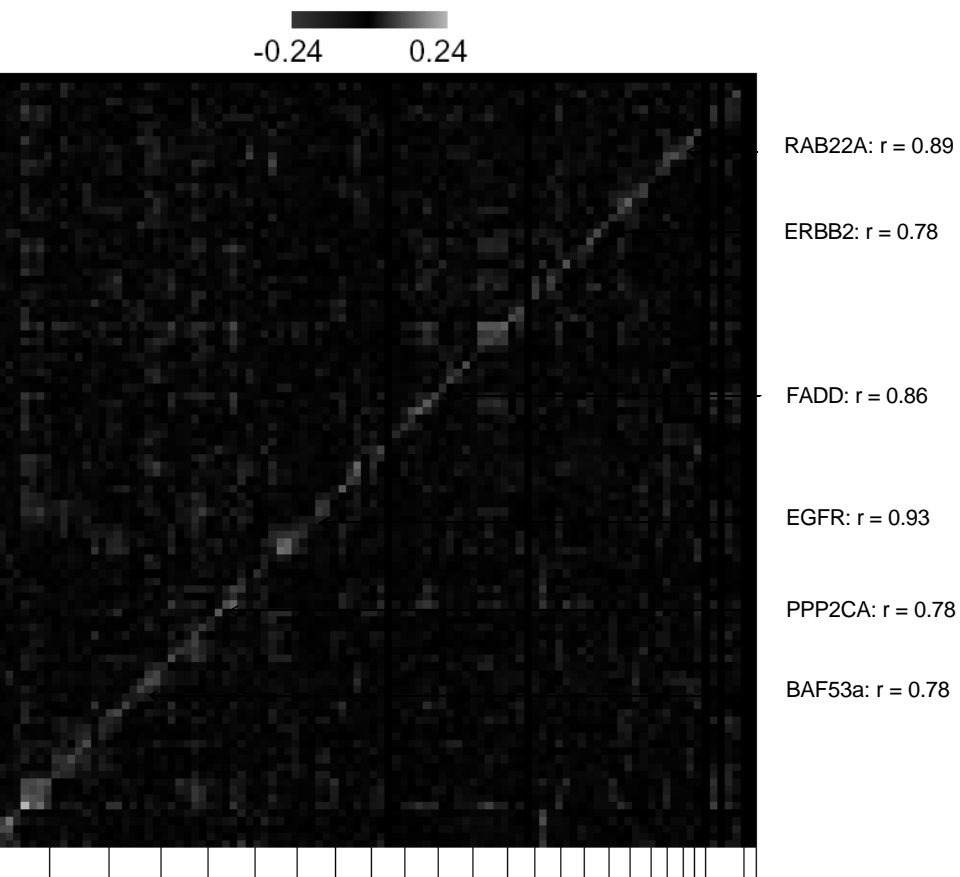
070124D1D32342EE86CE18FE3289A40

Ready NUM

44 Breast cancer cell lines were analyzed for mRNA expression (Affy) and array based CGH

- ◆ Question: what is the effect of genomic copy number on gene expression?
- ◆ Look at sample to sample correlation of CGH/expression data, but bin by genomic position.
 - Look for genes whose expression correlates with copy number in frequently altered regions

Application of Magellan to Breast Cancer Cell Line Data



Genome Wide Correlation Plot

- ◆ There is a positive correlation between copy number and expression. Those genes that correlate strongly can be investigated further

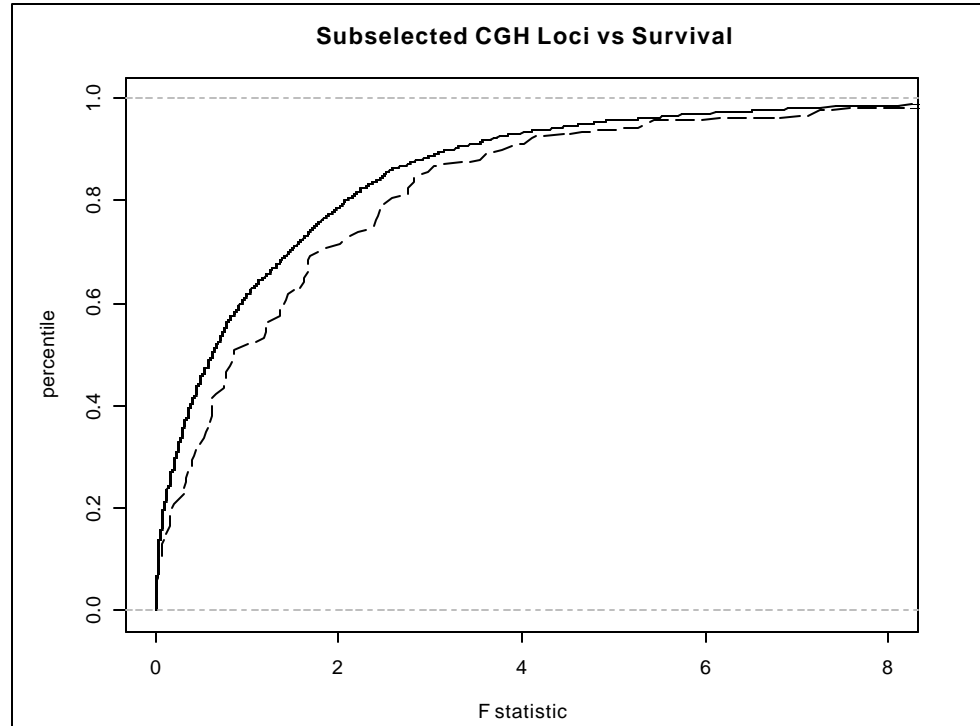
Projection of subsets between data types

- ◆ Looked at CGH and mRNA expression in 20 Ovarian tumor samples (10 long, 10 short survivors)
- ◆ Used curated annotations to find ‘equivalent’ variables from one data type to another
 - Annotations can be used as a means of establishing variable equivalence
 - Equivalence is user defined (string equality, numerical comparisons, etc).

Data → Identifiers → Annotations ↔ Annotations ← Identifiers ← Data

Projection of subsets between data types

- ◆ If we select for genes whose mRNA expression correlates with an outcome, do copy number changes of loci that map close to those genes also correlate?
 - Select Genes that correlate with patient survival
 - Project those genes onto CGH space – select those loci that map within 1Mb of the genes
 - Look at the correlation values of the sub selected loci vs. randomly chosen loci
- ◆ This sequence of tasks can be broken down into a series of simple operations in Magellan
 - Correlate expression with survival – store as a quantitative annotation
 - Sub select expression data
 - Project onto CGH data
 - Correlate the sub selected CGH loci with survival
 - Plot the results



Correlation of Sub Selected Loci vs. Patient Survival

- ◆ CGH loci located in close genomic proximity to genes that correlate with survival correlate better with survival than loci chosen at random ($p < 0.05$).

Magellan allows researchers to perform visualizations and analyses of their data in a web based environment

- ◆ Abstract representation of data and annotations insures a broad applicability
- ◆ Subsetting functionality allows users to sub select data based on qualitative and quantitative annotations
 - Useful for the creation of biologically meaningful sub sets as well as a means of reducing the effects of multiple comparisons
- ◆ Analytical methods can be deployed in a modular fashion
- ◆ Generalized methods can be combined to facilitate complex analyses
 - Sub selection, projection, visualization, import, export, etc.

Deliverables for caBIG:

- ◆ Interoperability of Magellan with caArray and caBIO
 - UML modeling of objects
 - Accessing information (especially curated annotations) from caArray
 - Decisions on use of / interface with existing caBIO objects.

Education of End Users

- ◆ Statistics shouldn't be a total 'black box' to experimentalists who are using tools like these

Experimental collaborators

- ◆ Gray Lab
 - Daniel Pollikof, Wen-Lin Kuo
- ◆ McCormick Lab
 - Jennifer Yeh
- ◆ Andy Berchuck (Duke)

Jain Lab

- ◆ Jane Fridlyand, PhD
- ◆ Lawrence Hon
- ◆ Barbara Novak
- ◆ Adam Olshen, PhD
- ◆ Tuan Pham
- ◆ Taku Tokuyasu, PhD